Motion-aware Contrastive Learning for Temporal Panoptic Scene Graph Generation

Thong Thanh Nguyen[1], **Xiaobao Wu**[2], **Yi Bin**[1,3*], **Cong-Duy T Nguyen**[2], **See-Kiong Ng**[1], **Anh Tuan Luu**[2]

[1] Institute of Data Science (IDS), National University of Singapore, Singapore
[2] Nanyang Technological University (NTU), Singapore,
[3] Tongji University, China

Abstract

To equip artificial intelligence with a comprehensive understanding towards a temporal world, video and 4D panoptic scene graph generation abstracts visual data into nodes to represent entities and edges to capture temporal relations. Existing methods encode entity masks tracked across temporal dimensions (mask tubes), then predict their relations with temporal pooling operation, which does not fully utilize the motion indicative of the entities' relation. To overcome this limitation, we introduce a contrastive representation learning framework that focuses on motion pattern for temporal scene graph generation. Firstly, our framework encourages the model to learn close representations for mask tubes of similar subject-relation-object triplets. Secondly, we seek to push apart mask tubes from their temporally shuffled versions. Moreover, we also learn distant representations for mask tubes belonging to the same video but different triplets. Extensive experiments show that our motion-aware contrastive framework significantly improves state-of-the-art methods on both video and 4D datasets.

Introduction

The advent of autonomous agents, intelligent systems, and robots warrants a comprehensive understanding of real-world environments (Ma et al. 2022; Driess et al. 2023; Raychaudhuri et al. 2023; Cheng et al. 2022; Li et al. 2023b,a). Such understanding encompasses beyond merely recognizing individual entities, but also a sophisticated understanding of their relationships. To construct a detailed understanding, scene graph generation (SGG) research (Li, Yang, and Xu 2022; Bin et al. 2019; Sudhakaran et al. 2023; Nag et al. 2023; Wang et al. 2024) has sought to provide relational perspective on scene understanding. In SGG frameworks, scene graphs utilize nodes to represent entities and edges to represent relationships, constructing a comprehensive and structured understanding of visual scenes.

However, due to being primarily based on bounding boxes to denote entities, scene graphs fall short of replicating human visual perception with a lack of granularity (Yang et al. 2023). To overcome this limitation, panoptic scene graph generation (Yang et al. 2022; Zhao et al. 2023) has been presented to expand the scope of SGG to incorporate pixel-level precise entity localization and thorough scene understanding including background components. Subsequently, because the temporal dimension undoubtedly provides richer information than the static spatial dimension, recent works (Yang et al. 2023, 2024) have shifted attention to the domain of videos and 4D scenes, resulting in the tasks of panoptic video and 4D scene graph generation.

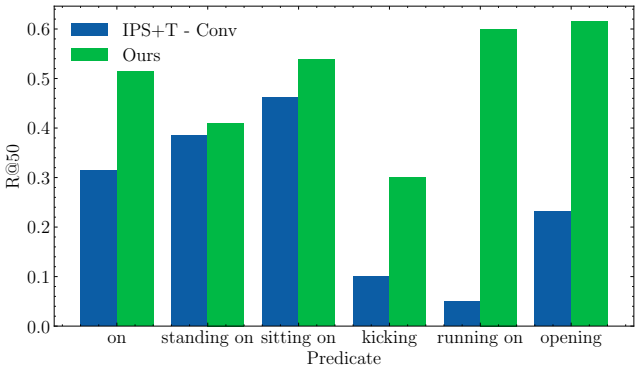

Figure 1: State-of-the-art IPS+T - Convolution (Yang et al. 2023) exhibits high R@50 scores for static relations, *e.g. on*, *sitting on*, and *standing on*, than dynamic relations, *e.g. kicking*, *running on*, and *opening*. In contrast, our method can perform effectively on both static and dynamic relations.

Popular methods (Yang et al. 2023, 2024, 2022) for temporal panoptic scene graph generation produce entity masks tracked across the temporal dimension, *i.e.* mask tubes, then predict temporal relations among them. To conduct relation prediction, these methods encode the segmentation mask tubes, apply global pooling, then forward to a multi-layer perceptron for classifying their relations. However, such global pooling operation is well-known to be ineffective in representing temporal or motion patterns, which are useful for determining the relation among the entities. Consequently, this would result in higher misclassification rates of more dynamic relations (Wang et al. 2023; Nag et al. 2023; Zhou et al. 2022), as illustrated in Figure 1.

To encourage temporal representation learning, current research (Nguyen et al. 2023; Liu et al. 2022; Zhou, Liu, and Wang 2023) uses contrastive learning for videos. However, they mainly seek to force two clips from the same video to be close together. As such, they mostly capture the semantics

*Yi Bin is the corresponding author, yi.bin@hotmail.com

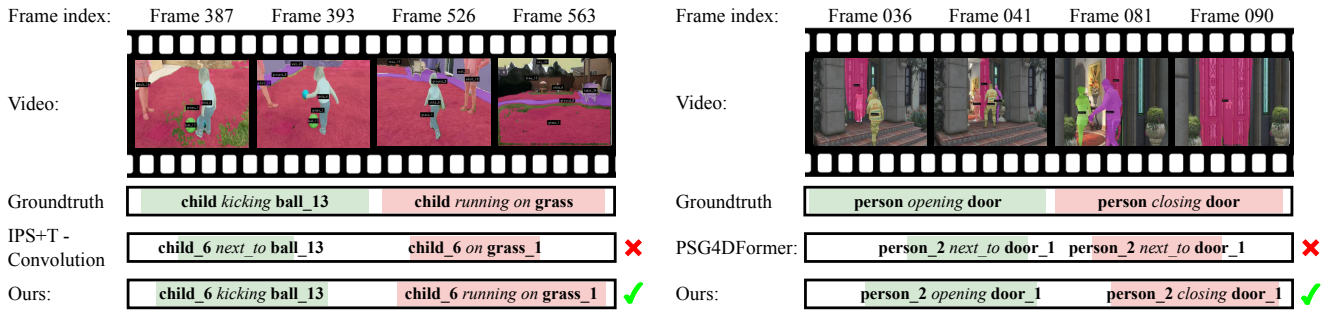

Figure 2: Examples of temporal panoptic scene graph generation of state-of-the-art (Yang et al. 2023, 2024) and our method.

of visual scenes and disregard motions (Chen et al. 2020). Moreover, instead of paying attention to precise entity localization, they work on frame-level representations. This would inadvertently inject motions from non-target entities into visual representations, which might not benefit relation classification in panoptic scene graph generation.

In this paper, to encourage representation learning to capture motion patterns for temporal panoptic scene graph generation, we propose a novel contrastive learning framework that focuses on mask tubes of the segmented entities. First, we force a mask tube and the one of similar subject-relation-object but of a different video to obtain close representations. Since positive mask tubes originate from distinct video clips, the model cannot rely upon visual semantics to optimize the contrastive objective, but instead depends on the motion trajectory evolution, which is our target component for representation learning. Second, we propel negative mask tubes generated by temporally shuffling the original tubes. Moreover, we also push apart representations of mask tubes from the same video but belonging to different triplets. Because mask tubes of different triplets from a common video with close visual features are separated from each other, we once again motivate the model to generate representations that are less reliant upon visual semantics but motion-sensitive features. In addition, the visually similar negative mask tubes can play a role as hard negative samples, thus accelerating the contrastive learning process (Chen, Zheng, and Song 2024).

Moreover, in order to implement our motion-aware contrastive learning framework, there is a need to quantify the relationship between mask tubes. This quantification marks a challenging problem as mask tubes are a sequence of segmentation masks that span over the sequence of video frames. Furthermore, mask tubes of two triplets might exhibit different lengths since two events often occur at different speed. Unfortunately, the popular pipeline of temporal pooling and then similarity estimation flattens the temporal dimension of the mask tubes and neglects their motion features. To resolve this problem, we consider mask tubes of two triplets as two distributions and seek the optimal transportation map between them, then utilize the transport distance as the distance between two triplets' tubes. Such scheme of transporting can play a role of synchronizing the motion states of two triplets and takes advantage of the mask tubes' evolutionary trajectory.

To sum up, our contributions are as follows:

- We propose a novel contrastive learning framework for temporal panoptic scene graph generation which pulls together entity mask tubes with similar motion patterns and pushes away those of distinct motion patterns.
- We utilize optimal transport distance to estimate the relationship between two events' mask tubes for the proposed contrastive framework.
- Comprehensive experiments demonstrate that our framework outperforms state-of-the-art methods on both natural and 4D video datasets, especially on recognizing dynamic subject-object relations.

Problem Formulation

Temporal panoptic scene graph generation (TPSGG) is a task to generate a dynamic scene graph given an input video. In the generated scene graph, each node corresponds to an entity and each edge corresponds to a spatial-temporal relation between two entities. Formally, the input of a TPSGG model is a video clip V, particularly $V \in \mathbb{R}^{T\times H\times W\times 3}$ for a natural video, $V \in \mathbb{R}^{T\times H\times W\times 4}$ for a 4D RGB-D video, and $V \in \mathbb{R}^{T\times M\times 6}$ for a 4D point cloud video, T denotes the number of frames, M the number of point clouds of interest, and the frame size $H \times W$ should remain consistent across the video. The output of the model is a dynamic scene graph G. The TPSGG task can be formulated as follows:

$$P(G|V) = P(M, O, R|V). \tag{1}$$

In particular, G consists of binary mask tubes $M = \{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_N\}$ and entity labels $O = \{o_1, o_2, ..., o_N\}$ which are associated with N entities in the video, and their relations are denoted as $R = \{r_1, r_2, ..., r_L\}$. With respect to entity o_i, the mask tube $\mathbf{m}_i \in \{0, 1\}^{T\times H\times W}$ composes all tracked masks in all video frames, and its category $o_i \in \mathbb{C}^O$. For all entities in frame t, their masks must not overlap, *i.e.* $\sum_{i=1}^{N} \mathbf{m}_i^t \leq \mathbf{1}^{H\times W}$. The relation $r_i \in \mathbb{C}^R$ associates two entities, one of which is the subject and the other is an object, with a relation class and a time period. $\mathbb{C}^O$ and $\mathbb{C}^R$ denote the entity and relation set class, respectively.

Methodology

We firstly present the backbone pipeline to conduct temporal panoptic scene graph generation. Then, we explain our proposed contrastive learning framework to facilitate motion-aware mask tube representation learning. We also present

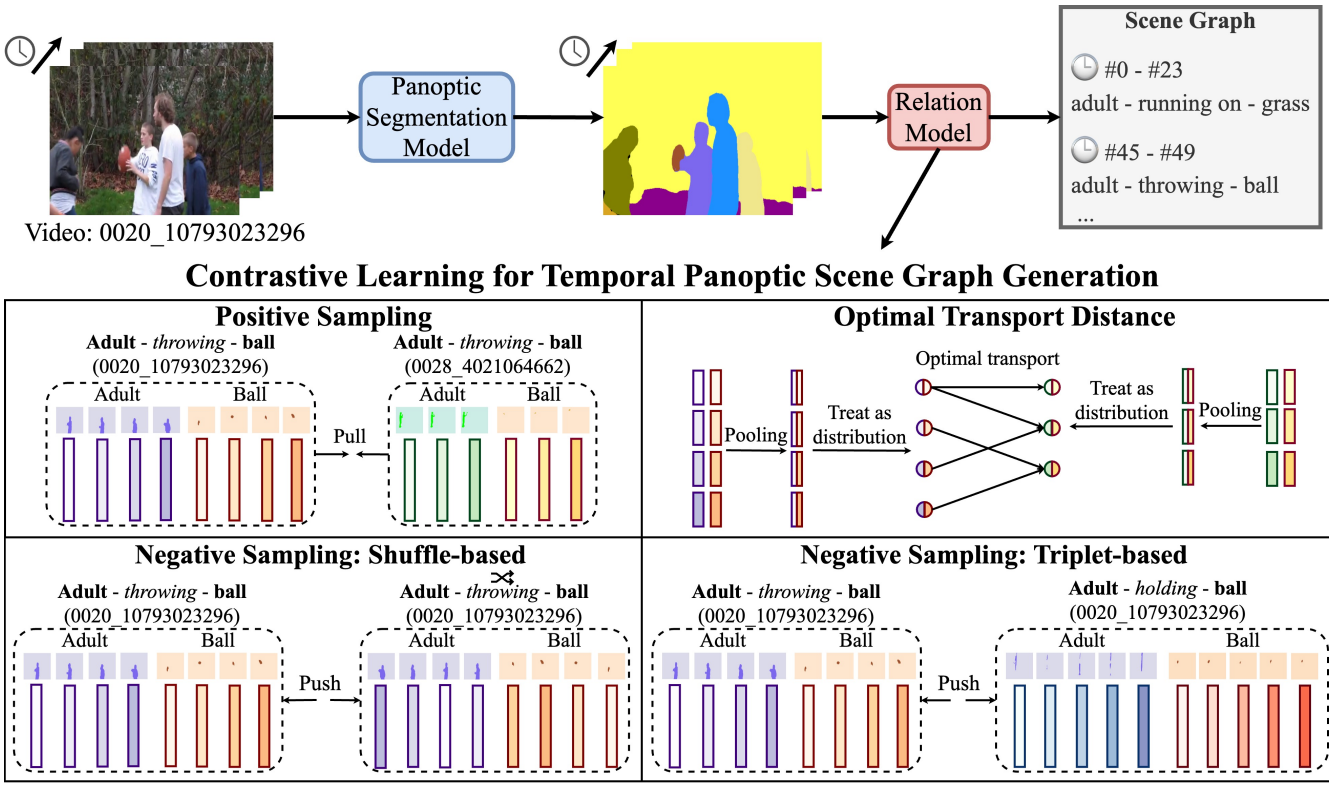

Figure 3: Framework overview of contrastive learning for temporal scene graph generation.

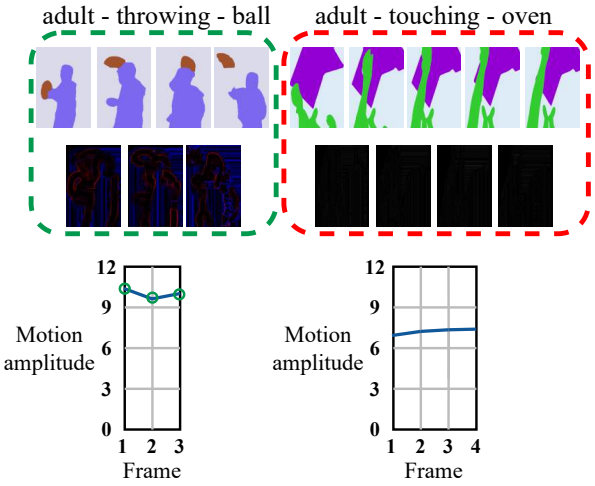

Figure 4: Proposed strategy to select strong-motion tubes.

the detail of our optimal transport approach to estimate the relation between two mask tubes for the contrastive objective. Our overall framework is illustrated in Figure 3.

Temporal Panoptic Segmentation

Given a video clip $V \in \mathbb{R}^{T\times H\times W\times 3}$, $V \in \mathbb{R}^{T\times H\times W\times 4}$, or $V \in \mathbb{R}^{T\times M\times 6}$, the initial step is to segment and track each pixel in a non-overlapping manner. Formally, the model produces a set of entity masks $\{y_i\}_{i=1}^{N} = \{(\mathbf{m}_i, p_i(c))\}_{i=1}^{N}$, where $\mathbf{m}_i \in \{0,1\}^{T\times H\times W}$ denotes the tracked video mask, *i.e.* the mask tube, and $p_i(c)$ denotes the probability of assigning class c to the tube $\mathbf{m}_i$. N denotes the number of entities, which consist of both foreground (thing) and background (stuff) classes.

Segmentation module. Inspired by (Yang et al. 2023, 2024), we adopt the Transformer-based encoder-decoder segmentation model. There are two types of segmentation procedure: 1) image panoptic segmentation combined with a tracker (IPS+T) and 2) video panoptic segmentation (VPS). IPS+T procedure will process each video frame separately and uses the tracker to connect the mask tubes across the video frames, while VPS processes each video frame with its reference frame from a nearby timestamp.

Both procedures are initiated by producing a set of object queries which interacts with encoded visual patches via masked cross-attention. Receiving a video V, the model produces a set of queries $\{\mathbf{q}_i\}_{i=1}^{N}$, where each query $\mathbf{q}_i$ corresponds to one entity. Subsequently, every query is forwarded to two multi-layer perceptrons (MLPs) to project the queries into mask classification and mask regression outputs.

Training and inference. During training, each query is matched to a groundtruth mask through mask-based bipartite matching to calculate the segmentation loss. During inference, IPS+T generates panoptic segmentation masks for each frame, and uses the tracker to achieve N tracked mask tubes. In contrast, VPS employs two query embeddings of the target and reference frame, and performs query-wised similarity tracking to obtain N tracked mask tubes.

Relation Classification

After the segmentation step, if the relation module is to be trained, we match query tubes with the annotated groundtruth masks based on the tube IoU values with the groundtruth. Otherwise, we directly forward mask tubes to self-attention or convolutional layers for encoding them into hidden representations $\{H_i\}_{i=1}^{N}$, $H_i \in \mathbb{R}^{T\times D}$, where D denotes the hidden dimension. Then, we construct query pairs from every two query tubes' representations H_i and H_j, $i, j \in \{1, 2, ..., N\},\ i \neq j$. Inspired by (Yang et al. 2023, 2024), in every pair, we perform global pooling over the temporal dimension for each mask tube representation:

$$\mathbf{h}_i = \text{Pooling}\left(H_i\right), \tag{2}$$

where $\mathbf{h}_i \in \mathbb{R}^D$. Afterwards, we concatenate $\mathbf{h}_i$ and $\mathbf{h}_j$, and forward to a MLP to generate the relation category:

$$\log p(r_{i,j}) = \text{MLP}\left([\mathbf{h}_i, \mathbf{h}_j]\right). \tag{3}$$

To train the relation classification module, we use the cross-entropy loss calculated based on the predicted relation log-likelihood and the groundtruth. For inference, we extract the relation of the highest log-likelihood.

Contrastive Learning for Temporal Panoptic Scene Graph Generation

Our goal is to encourage mask tube representations $\{H_i\}_{i=1}^{N}$ to become motion-aware. In the beginning, we concatenate the representations of two mask tubes H_i^{sub} and H_j^{obj}, which have been matched to a groundtruth subject-relation-object triplet, to form an anchor representation $H_{i,j}$ (anchor):

$$H_{i,j}^{a} = [H_i^{\text{sub}}, H_j^{\text{obj}}], \tag{4}$$

where $H_{i,j}^{a} \in \mathbb{R}^{T\times 2D}$. Then, we propose a contrastive learning framework in which we motivate the model to associate mask tubes based upon the motion information. The objective of contrastive learning is to produce a representation space through attracting positive pairs, *i.e.* H^a and H^p (positive), while pushing apart negative pairs, *i.e.* H^a and H^n (negative). We accomplish this by optimizing the contrastive objective, which is formulated as follows:

$$\mathcal{L}_{\text{cont}} = -\log \frac{e^{\text{sim}(H^a, H^p)}}{e^{\text{sim}(H^a, H^p)} + \sum_{z=1}^{N_n} e^{\text{sim}(H^a, H_z^n)}}, \tag{5}$$

where sim denotes the similarity function defined upon a pair of mask tube representations. The formulation shows that what the model learn is largely dependent upon how positive and negative samples are generated.

Positive sampling. To satisfy our motion-aware requirement for contrastive learning, we extract mask tube representations from the entities of the same subject and object category that exhibit a similar groundtruth relation from another video. Since two videos possess distinct visual features, the model must rely on the shared motion pattern of similar subject-relation-object triplets to associate the anchor and the positive sample.

Negative sampling. For negative sampling, we design two strategies, which result in two contrastive approaches, *i.e.* shuffle-based and triplet-based contrastive learning.

Shuffle-based contrastive learning

In our shuffle-based approach, we create negative samples by utilizing a series of temporal permutations π to the anchor tube, *i.e.* shuffling:

$$H^n = \pi\left(H^a\right). \tag{6}$$

As such, the contrastive objective will force the model to propel representations of the anchor tube, which is in the normal order, from the shuffled tube, which exhibits a distorted motion due to the shuffled order. This would make the learned representation sensitive to frame ordering, *i.e.* motion-aware, as the anchor H^a and the negative tube H^n share visual semantics and can only be distinguished using motion information.

Selecting strong-motion mask tubes. However, there exists a potential risk: for static relations such as *on*, *next to*, and *in*, mask tubes might involve almost no motion. As a result, the shuffled tube would become identical to the anchor one and the model would not be able to differentiate them and learn reasonably. To address this problem, we propose a strategy to select strong-motion tubes for shuffling, which we illustrated in Figure 4.

Given a video, our aim is to select mask tubes that carry strong motion for shuffling. To measure the motion of the mask tube, we utilize optical flow edges (Xiao, Tighe, and Modolo 2021). We estimate flow edges via employing a Sobel filter (Sobel et al. 2022) onto the flow magnitude map and take the median over the flow edge pixels of the entity masks. Then, we select mask tubes whose the maximum value across the optical flow surpasses a threshold γ.

Triplet-based contrastive learning

To take advantage of motion-aware signals from triplets of similar subject-relation-object category, we design a triplet-based approach to create negative samples. A naive approach would be to sample mask tubes of any distinct subject-relation-object triplet from the anchor sample. However, if we run into triplets with all distincgt subject, relation, and object categories, the negative pair would be trivial for the model to distinguish, resulting in less effective learning.

In order to create harder negative samples, we choose negative mask tubes from the same video with the anchor. We create a multi-nomial distribution, where triplets that share more subject, relation, or object categories with the anchor will be more likely to be drawn. Hence, our negative samples can hold close visual semantics with the anchor sample, and increase the likelihood that the model depends on motion semantics to push them apart. From contrastive learning perspective, these samples form hard negative samples to accelerate the learning process (Chen, Zheng, and Song 2024).

Optimal Transport for Mask Tube Relation Quantification

There is one remaining problem, *i.e.* how to define the similarity function sim for two mask tubes' representations H_i and H_j. In this work, we consider two mask tubes as two discrete distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, whose H_i and H_j are

Algorithm 1: Computing the optimal transport distance

Require: $\mathbf{C} = \{\mathbf{C}_{l,k} = c(\mathbf{h}_{i,l}, \mathbf{h}_{j,k}) \mid 1 \le i \le N_V, 1 \le j \le N_L\} \in \mathbb{R}^{T_i \times T_j}$, $\mathbf{a} \in \mathbb{R}^{T_i}$, $\mathbf{b} \in \mathbb{R}^{T_j}$, s, N_{iter}

$d_{\text{OT}} = \infty$

for $s = 1$ to $\min(T_i, T_j)$ **do**

$\mathbf{T} = \exp\left(-\frac{\mathbf{C}}{\tau}\right)$

$\mathbf{T} = \frac{s}{(\mathbf{1}_{T_i})^\top \cdot \mathbf{T} \cdot \mathbf{1}_{T_j}} \mathbf{T}$

for $i = 1$ to N_{iter} **do**

$\mathbf{p}_a = \min\left(\frac{\mathbf{a}}{\mathbf{T}\mathbf{1}_{T_j}}, \mathbf{1}_{T_i}\right)$, $\mathbf{T}_a = \text{diag}(\mathbf{p}_a) \cdot \mathbf{T}$

$\mathbf{p}_b = \min\left(\frac{\mathbf{b}}{\mathbf{T}_a^\top \mathbf{1}_{T_i}}, \mathbf{1}_{T_j}\right)$, $\mathbf{T}_b = \text{diag}(\mathbf{p}_b) \cdot \mathbf{T}_a$

$\mathbf{T} = \frac{s}{(\mathbf{1}_{T_i})^\top \cdot \mathbf{T} \cdot \mathbf{1}_{N_L}} \mathbf{T}_b$

end for

$d_{\text{OT}} = \min\left(d_{\text{OT}}, \sum_{k=1}^{T_i} \sum_{l=1}^{T_j} \mathbf{T}_{k,l} \mathbf{C}_{k,l}\right)$

end for

return d_{OT}

their supports, respectively. Formally, $\boldsymbol{\mu} = \sum_{k=1}^{T_i} \mathbf{a}_k \delta_{\mathbf{h}_{i,k}}$ and $\boldsymbol{\nu} = \sum_{l=1}^{T_j} \mathbf{b}_l \delta_{\mathbf{h}_{j,l}}$, where $\delta_{\mathbf{h}_{i,k}}$ and $\delta_{\mathbf{h}_{j,l}}$ denote the Dirac functions centered upon $\mathbf{h}_{i,k}$ and $\mathbf{h}_{j,l}$, respectively. The weights of the supports are $\mathbf{a} = \frac{\mathbf{1}_{T_i}}{T_i}$ and $\mathbf{b} = \frac{1_{T_j}}{T_j}$.

After defining the distribution scheme, we propose the tube alignment optimization problem, which is to find the transport plan that achieves the minimum distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as follows:

$$d_{\text{OT}} = \mathcal{D}_{\text{OT}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{k=1}^{T_i} \sum_{l=1}^{T_j} \mathbf{T}_{i,j} \cdot c(\mathbf{h}_{i,k}, \mathbf{h}_{j,l}), \tag{7}$$

$$\text{s.t} \quad \Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{T} \in \mathbb{R}_+^{T_i \times T_j} \mid \mathbf{T}\mathbf{1}_{T_i} \le \mathbf{a}, \mathbf{T}^\top \mathbf{1}_{T_j} \le \mathbf{b}, \mathbf{1}_{T_i}^\top \cdot \mathbf{T} \cdot \mathbf{1}_{T_j} = s, \quad 0 \le s \le \min(T_i, T_j)\}, \tag{8}$$

where c denotes a pre-defined distance between two vectors. We implement the cost distance $c(\mathbf{h}_{i,k}, \mathbf{h}_{j,l}) = 1 - \frac{\mathbf{h}_{i,k} \cdot \mathbf{h}_{j,l}}{||\mathbf{h}_{i,k}||_2 ||\mathbf{h}_{i,k}||_2}$ as the cosine distance. As the exact optimization over the transport plan $\mathbf{T}$ is intractable, we adopt the Sinkhorn-based algorithm to estimate $\mathbf{T}$. We delineate the algorithm to calculate the distance in Algorithm 1. To turn the distance into similarity value, we take its negative value and add to a pre-defined margin α:

$$\text{sim}(\mathbf{h}^a, \mathbf{h}^p) = \alpha - d_{\text{OT}}. \tag{9}$$

Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our motion-aware contrastive framework. We first describe the experiment settings, covering the evaluation datasets, evaluation metrics, baseline methods, and implementation details. Next, we present quantitative results of our method, then provide ablation study and careful analysis to explore properties of our motion-aware contrastive framework. Eventually, we conduct qualitative analysis to concretely examine its behavior.

Experiment Settings

Datasets. We assess the effectiveness of our method on natural and 4D video inputs. The corresponding dataset to each input type is as follows:

- **Open-domain Panoptic video scene graph generation (OpenPVSG)** (Yang et al. 2023): OpenPVSG consists of scene graphs and associated segmentation masks with respect to subject and object nodes in the scene graph. The dataset comprises 400 videos, including 289 third-person videos from ViDOR (Shang et al. 2019), 111 egocentric videos from Epic-Kitchens (Damen et al. 2022) and Ego4D (Grauman et al. 2022).
- **Panoptic scene graph generation for 4D (PSG4D)** (Yang et al. 2024): The PSG4D dataset is divided into two groups, *i.e.* PSG4D-GTA and PSG4D-HOI. PSG4D-GTA comprises 67 third-view videos with an average length of 84 seconds, 35 object categories, and 43 relationship categories. On the contrary, PSG4D-HOI contains 2,973 videos from an egocentric perspective, whose average duration is 20 seconds. The PSG4D-HOI's videos are mostly related to indoor scenes, covering 46 object categories and 15 relationship categories.

Evaluation metrics. We use the recall at K (R@K) and mean recall at K (mR@K) metrics, which are standard metrics used in scene graph generation tasks. Both R@K and mR@K consider the top-K triplets predicted by the panoptic scene graph generation model. A successful recall of a predicted triplet must satisfy the following criteria: 1) correct category labels for the subject, object, and predicate; 2) a volume Intersection over Union (vIoU) greater than or equal to 0.5 between the predicted mask tubes and the groundtruth tubes. For extensive comparison, we also report results with the vIoU threshold of 0.1.

Baseline methods. We compare our method with a comprehensive list of baseline approaches for temporal panoptic scene graph generation: (i) **IPS+T - Vanilla** (Yang et al. 2023) uses image panoptic segmentation (IPS) model with a tracker for segmentation, and fully-connected layers to separately encode temporal states of entity mask tubes; (ii) **IPS+T - Handcrafted filter** (Yang et al. 2023) uses image panoptic segmentation (IPS) model with a tracker for segmentation, and a manually-designed kernel to encode entity mask tubes; (iii) **IPS+T - Convolution** (Yang et al. 2023) uses image panoptic segmentation (IPS) model with a tracker for segmentation, and learnable convolutional layers to encode entity mask tubes; (iv) **IPS+T - Transformer** (Yang et al. 2023) uses image panoptic segmentation model (IPS) with a tracker for segmentation, and Transformer-based encoder with self-attention layers to encode entity mask tubes; (v) **VPS - Vanilla** (Yang et al. 2023) is similar to IPS+T - Vanilla, but uses video panoptic segmentation (VPS) model for panoptic segmentation; (vi) **VPS - Handcrafted filter** (Yang et al. 2023) is similar to IPS+T - Handcrafter filter, but uses video panoptic segmentation (VPS) model for segmentation; (vii) **VPS - Convolution** (Yang et al. 2023) is similar to IPS+T - Convolution, but uses video panoptic segmentation (VPS) model for segmentation; (viii) **VPS - Transformer** (Yang et al. 2023) is similar to IPS+T

Method	vIoU threshold = 0.5			vIoU threshold = 0.1		
	R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
IPS+T - Vanilla	3.04 / 1.35	4.61 / 2.94	5.56 / 3.33	8.28 / 5.68	14.47 / 9.92	18.24 / 11.84
IPS+T - Handcrafted filter	2.52 / 1.72	3.77 / 2.36	4.72 / 2.79	8.07 / 5.61	13.42 / 8.27	16.46 / 10.11
IPS+T - Transformer	3.88 / 2.81	5.66 / 4.12	6.18 / 4.44	9.01 / 6.69	14.88 / 11.28	17.51 / 13.20
IPS+T - Convolution	3.88 / 2.55	5.24 / 3.29	6.71 / 5.36	10.06 / 8.98	14.99 / 12.21	18.13 / 15.47
Ours - Transformer	3.98 / 2.98	5.97 / 4.20	7.44 / 5.15	10.59 / 9.56	16.98 / 12.39	22.33 / 17.47
Ours - Convolution	**4.51 / 3.56**	**6.08 / 4.38**	**7.76 / 5.86**	**11.43 / 9.57**	**17.30 / 13.13**	**22.85 / 17.48**
VPS - Vanilla	0.21 / 0.10	0.21 / 0.10	0.31 / 0.18	6.29 / 3.04	9.64 / 6.74	12.89 / 9.60
VPS - Handcrafted filter	0.42 / 0.13	0.52 / 0.50	0.94 / 0.92	5.24 / 2.84	7.65 / 7.14	9.64 / 8.22
VPS - Transformer	0.42 / 0.61	0.73 / 0.76	1.05 / 0.92	6.50 / 5.75	9.64 / 8.25	12.26 / 9.51
VPS - Convolution	0.42 / 0.25	0.63 / 0.67	0.63 / 0.67	8.07 / 7.84	11.01 / 9.78	12.89 / 10.77
Ours - Transformer	0.63 / 0.83	1.05 / 0.76	1.05 / 0.76	6.71 / 6.94	10.27 / 8.68	13.42 / 12.09
Ours - Convolution	**0.84 / 0.98**	**1.26 / 1.22**	**1.26 / 1.22**	**8.18 / 8.00**	**12.90 / 11.47**	**14.22 / 13.59**

Table 1: Experimental results on the OpenPVSG dataset.

Input type	Method	PSG4D-GTA			PSG4D-HOI		
		R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
Point cloud videos	3DSGG	1.48 / 0.73	2.16 / 0.79	2.92 / 0.85	3.46 / 2.19	3.15 / 2.47	4.96 / 2.84
	PSG4DFormer	4.33 / 2.10	4.83 / 2.93	5.22 / 3.13	5.36 / 3.10	5.61 / 3.95	6.76 / 4.17
	Ours	**5.88 / 3.45**	**6.31 / 3.70**	**7.31 / 4.70**	**7.28 / 5.09**	**7.62 / 6.49**	**9.18 / 6.85**
RGB-D videos	3DSGG	2.29 / 0.92	2.46 / 1.01	3.81 / 1.45	4.23 / 2.19	4.47 / 2.31	4.86 / 2.41
	PSG4DFormer	6.68 / 3.31	7.17 / 3.85	7.22 / 4.02	5.62 / 3.65	6.16 / 4.16	6.28 / 4.97
	Ours	**9.07 / 5.52**	**9.73 / 6.32**	**9.73 / 6.32**	**7.63 / 6.09**	**8.36 / 6.94**	**8.53 / 8.29**

Table 2: Experimental results on both PSG4D-GTA and PSG4D-HOI groups of PSG4D dataset.

Method	R/mR@20	R/mR@50	R/mR@100
w/o shuffle-based	4.41 / 3.43	5.90 / 4.24	7.30 / 5.79
w/o triplet-based	4.44 / 3.50	6.02 / 4.28	7.36 / 5.83
Ours	**4.51 / 3.56**	**6.08 / 4.38**	**7.44 / 5.86**

Table 3: Ablation results for contrastive learning approaches on OpenPVSG dataset. We adopt the vIoU threshold of 0.5.

Tube relation quantification	R/mR@20	R/mR@50	R/mR@100
Pooling - Cosine similarity	4.44 / 3.40	6.04 / 4.36	7.36 / 5.84
Pooling - L2	4.36 / 3.37	3.77 / 5.95	7.29 / 5.80
Optimal transport	**4.51 / 3.56**	**6.08 / 4.38**	**7.44 / 5.86**

Table 4: Ablation results for mask tube relation quantification method between mask tubes on OpenPVSG dataset.

- Transformer, but uses video panoptic segmentation (VPS) model for segmentation; (ix) **3D-SGG** (Wald et al. 2020) is based on PointNet (Qi et al. 2017) and graph convolutional network (Kipf and Welling 2016) but neglects the depth dimension and generates panoptic scene graphs for 4D video inputs; (x) **PSG4DFormer** (Yang et al. 2024) is a specialized model for 4D inputs, using Mask2Former (Cheng et al. 2022) for segmentation and a spatial-temporal Transformer to encode object mask tubes for relation classification.

Implementation details. For fair comparison, we experiment our contrastive framework with both IPS+T and VPS as segmentation module for panoptic video scene graph generation. In the former case, we leverage the UniTrack tracker (Wang et al. 2021) combined with Mask2Former model (Cheng et al. 2022), which is initialized from the best-performing COCO-pretrained weights and fine-tuned for 8 epochs using AdamW optimizer with a batch size of 32, learning rate of 0.0001, weight decay of 0.05, and gradient clipping with a max L2 norm of 0.01. In the latter case, we utilize Video K-Net (Li et al. 2022), also initialized from COCO-pretrained weights and fine-tuned with the same strategy as IPS+T. In the relation classification step, we conduct fine-tuning with a batch size of 32, employing the Adam optimizer with a learning rate of 0.001. For 4D panoptic scene graph generation, we adopt the PSG4DFormer baseline. To work with RGB-D and point cloud videos, we use an ImageNet pretrained on ResNet-101 (Russakovsky et al. 2015) and the DKNet (Wu et al. 2022) as the visual encoder, respectively. We fine-tune the segmentation module for RGB-D and point cloud videos for 12 and 200 epochs, respectively. We use additional 100 epochs to train the relation classification module. Based on validation, we adopt a threshold $\gamma = 9.0$ and a margin $\alpha = 10.0$. We set the maximum number of iterations N_{iter} to 1,000.

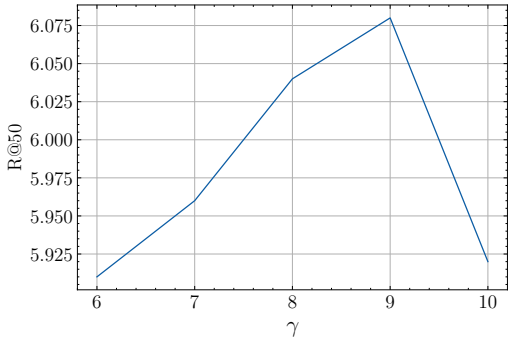

Figure 5: Ablation results on threshold γ.

Main Results

Results on OpenPVSG. As shown in Table 1, we substantially outperform both IPS+T - Convolution and IPS+T - Transformer when we use IPS+T for segmentation. In particular, using a higher vIoU threshold to filter out inaccurate segmentation, we surpass IPS+T - Transformer by 1.3/0.7

Input type	Method	PSG4D-GTA			PSG4D-HOI		
		R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
Point cloud videos	w/o shuffle-based	5.56 / 2.92	5.57 / 2.98	6.51 / 4.36	6.56 / 4.29	6.98 / 6.25	8.76 / 6.43
	w/o triplet-based	5.77 / 2.93	5.59 / 3.26	6.53 / 4.39	6.67 / 4.85	7.52 / 6.31	8.84 / 6.43
	Ours	**5.88 / 3.45**	**6.31 / 3.70**	**7.31 / 4.70**	**7.28 / 5.09**	**7.62 / 6.49**	**9.18 / 6.85**
RGB-D videos	w/o shuffle-based	8.35 / 5.34	8.76 / 5.68	8.88 / 5.53	7.00 / 5.53	7.51 / 6.02	7.56 / 7.42
	w/o triplet-based	9.00 / 5.46	9.71 / 5.95	9.63 / 5.82	7.12 / 6.03	8.31 / 6.51	8.24 / 7.95
	Ours	**9.07 / 5.52**	**9.73 / 6.32**	**9.73 / 6.32**	**7.63 / 6.09**	**8.36 / 6.94**	**8.53 / 8.29**

Table 5: Ablation results for contrastive learning approaches on PSG4D dataset.

Input type	Tube relation quantification	PSG4D-GTA			PSG4D-HOI		
		R/mR@20	R/mR@50	R/mR@100	R/mR@20	R/mR@50	R/mR@100
Point cloud videos	Pooling - Cosine similarity	5.76 / 2.87	6.02 / 3.62	6.84 / 4.11	7.24 / 4.45	7.44 / 6.27	8.20 / 6.64
	Pooling - L2	5.46 / 2.78	5.38 / 3.39	6.51 / 3.86	6.72 / 4.11	6.74 / 6.05	7.96 / 6.11
	Optimal transport	**5.88 / 3.45**	**6.31 / 3.70**	**7.31 / 4.70**	**7.28 / 5.09**	**7.62 / 6.49**	**9.18 / 6.85**
RGB-D videos	Pooling - Cosine similarity	9.03 / 5.37	9.47 / 5.86	9.70 / 6.02	7.36 / 5.43	7.93 / 6.70	8.06 / 7.42
	Pooling - L2	8.89 / 4.70	8.90 / 5.41	9.08 / 5.78	6.65 / 5.26	7.74 / 6.29	7.95 / 7.39
	Optimal transport	**9.07 / 5.52**	**9.73 / 6.32**	**9.73 / 6.32**	**7.63 / 6.09**	**8.36 / 6.94**	**8.53 / 8.29**

Table 6: Ablation results for mask tube relation quantification method between mask tubes on PSG4D dataset.

points of R/mR@100, while surpassing IPS+T - Convolution by 0.8/1.1 points of R/mR@50. In addition, for a less strict vIoU threshold, we outperform IPS+T - Transformer by 1.6/2.9 points of R/mR@20, and IPS+T - Convolution by 2.3/0.9 points of R/mR@50. These results demonstrate that our method makes a propitious contribution to temporal panoptic scene graph generation, not only to popular but also to unpopular relation classes.

Results on PSG4D. Table 2 shows that our method also achieves significantly higher performance than the PSG4DFormer model. Specifically, when working with point cloud videos, on PSG4D-GTA, we outperform the baseline method by 1.6/1.4 points. Analogously, on PSG4D-HOI, we outperform PSG4DFormer by 2.0/2.5 points of R/mR@50. These results indicate that our framework bears a valuable impact to both egocentric and third-view videos. We hypothesize that both video types consist of dynamic actions among objects whose mask tube representations should be polished. In addition, when working with RGB-D videos, on PSG4D-GTA, we enhance the baseline method by 2.4/2.2 points of R/mR@20,. Furthermore, on PSG4D-HOI, our motion-aware contrastive learning also considerably refines PSG4DFormer by 2.0/2.4 points of R/mR@20. Such results have verified the generalizability of our motion-aware contrastive framework over natural, point cloud, and RGB-D videos.

Ablation Study

Effect of the contrastive components. We evaluate our framework without the assistance of either the shuffle-based or the triplet-based contrastive objective. As shown in Table 3 and 5, the performance degrades when we both remove shuffle-based and triplet-based contrastive approaches. In addition, triplet-based contrastive learning plays a more fundamental role than the shuffle-based one. We hypothesize that shuffle-based contrastive learning is better at focusing on motion semantics than triplet-based one.

Effect of selecting strong-motion tubes. We evaluate the impact of our strategy to filter out weak motion tubes. In Figure 5, we observe a performance boost when we increase the threshold to select mask tubes with strong motion. However, further elevating the threshold results in performance degradation, since there are more mask tubes eliminated, thus limiting the effect of our motion-aware contrastive framework.

Effect of optimal transport distance. In this ablation, we compare various strategies to calculate the similarity between two mask tubes. Results in Table 4 and 6 show that the proposed optimal transport achieves much higher performance for both natural and 4D video inputs. We conjecture that other method such as pooling then cosine similarity or L2 neglects the temporal or flattens the motion nature of the entity mask tubes, thus reducing the effectiveness.

Qualitative Analysis

We visualize examples processed by the state-of-the-art models and ours in Figure 2. As can be observed, our model successfully produces mask tubes overlapping with the groundtruth, and importantly predicts the correct relations of the subject-object pairs. On the other hand, baseline models tend to prefer more static relations, since during training they do not explicitly focus on motion-sensitive features. Statistics in Figure 1 also substantiate our proposition, in which we achieve considerably higher recalls for dynamic relations than baseline approaches.

Conclusion

In this paper, we propose a motion-aware contrastive learning framework for temporal panoptic scene graph generation. In our framework, we learn close representations for temporal masks of similar entities that exhibit common relations. Moreover, we separate temporal masks from their shuffled version, and also separate temporal masks of different subject-relation-object triplets. To quantify the relationship among temporal masks in the proposed contrastive framework, we utilize optimal transport to preserve the temporal nature among temporal entity masks. Extensive experiments substantiate the effectiveness of our framework for both natural and 4D videos.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-051T). Thong Nguyen is supported by a Google Ph.D. Fellowship in Natural Language Processing.

References

Bin, Y.; Yang, Y.; Tao, C.; Huang, Z.; Li, J.; and Shen, H. T. 2019. Mr-net: Exploiting mutual relation for visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8110–8117.

Chen, Y.; Ma, G.; Yuan, C.; Li, B.; Zhang, H.; Wang, F.; and Hu, W. 2020. Graph convolutional network with structure pooling and joint-wise channel attention for action recognition. *Pattern Recognition*, 103: 107321.

Chen, Z.; Zheng, T.; and Song, M. 2024. Curriculum Negative Mining For Temporal Networks. *arXiv preprint arXiv:2407.17070*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2022. Epic-kitchens-100. *International Journal of Computer Vision*, 130: 33–55.

Driess, D.; Xia, F.; Sajjadi, M. S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, X.; Ding, H.; Yuan, H.; Zhang, W.; Pang, J.; Cheng, G.; Chen, K.; Liu, Z.; and Loy, C. C. 2023a. Transformer-based visual segmentation: A survey. *arXiv preprint arXiv:2304.09854*.

Li, X.; Yuan, H.; Zhang, W.; Cheng, G.; Pang, J.; and Loy, C. C. 2023b. Tube-Link: A flexible cross tube framework for universal video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13923–13933.

Li, X.; Zhang, W.; Pang, J.; Chen, K.; Cheng, G.; Tong, Y.; and Loy, C. C. 2022. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18847–18857.

Li, Y.; Yang, X.; and Xu, C. 2022. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13874–13883.

Liu, K.; Li, Y.; Xu, Y.; Liu, S.; and Liu, S. 2022. Spatial focus attention for fine-grained skeleton-based action tasks. *IEEE Signal Processing Letters*, 29: 1883–1887.

Ma, X.; Yong, S.; Zheng, Z.; Li, Q.; Liang, Y.; Zhu, S.-C.; and Huang, S. 2022. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*.

Nag, S.; Min, K.; Tripathi, S.; and Roy-Chowdhury, A. K. 2023. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22803–22813.

Nguyen, T.; Wu, X.; Dong, X.; Nguyen, C.-D.; Ng, S.-K.; and Tuan, L. A. 2023. Demaformer: Damped exponential moving average transformer with energy-based modeling for temporal language grounding. *arXiv preprint arXiv:2312.02549*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Raychaudhuri, S.; Campari, T.; Jain, U.; Savva, M.; and Chang, A. X. 2023. Reduce, reuse, recycle: Modular multi-object navigation. *arXiv preprint arXiv:2304.03696*, 2.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; and Chua, T.-S. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 279–287.

Sobel, I.; Duda, R.; Hart, P.; and Wiley, J. 2022. Sobel-feldman operator. *Preprint at https://www. researchgate. net/profile/Irwin-Sobel/publication/285159837. Accessed*, 20.

Sudhakaran, G.; Dhami, D. S.; Kersting, K.; and Roth, S. 2023. Vision relation transformer for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21882–21893.

Wald, J.; Dhamo, H.; Navab, N.; and Tombari, F. 2020. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3961–3970.

Wang, G.; Li, Z.; Chen, Q.; and Liu, Y. 2024. OED: Towards One-stage End-to-End Dynamic Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27938–27947.

Wang, W.; Luo, Y.; Chen, Z.; Jiang, T.; Yang, Y.; and Xiao, J. 2023. Taking a closer look at visual relation: Unbiased video scene graph generation with decoupled label learning. *IEEE Transactions on Multimedia*.

Wang, Z.; Zhao, H.; Li, Y.-L.; Wang, S.; Torr, P.; and Bertinetto, L. 2021. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34: 726–738.

Wu, Y.; Shi, M.; Du, S.; Lu, H.; Cao, Z.; and Zhong, W. 2022. 3d instances as 1d kernels. In *European Conference on Computer Vision*, 235–252. Springer.

Xiao, F.; Tighe, J.; and Modolo, D. 2021. Modist: Motion distillation for self-supervised video representation learning. *arXiv preprint arXiv:2106.09703*, 3.

Yang, J.; Ang, Y. Z.; Guo, Z.; Zhou, K.; Zhang, W.; and Liu, Z. 2022. Panoptic scene graph generation. In *European Conference on Computer Vision*, 178–196. Springer.

Yang, J.; Cen, J.; Peng, W.; Liu, S.; Hong, F.; Li, X.; Zhou, K.; Chen, Q.; and Liu, Z. 2024. 4d panoptic scene graph generation. *Advances in Neural Information Processing Systems*, 36.

Yang, J.; Peng, W.; Li, X.; Guo, Z.; Chen, L.; Li, B.; Ma, Z.; Zhou, K.; Zhang, W.; Loy, C. C.; et al. 2023. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18675–18685.

Zhao, C.; Shen, Y.; Chen, Z.; Ding, M.; and Gan, C. 2023. Textpsg: Panoptic scene graph generation from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2839–2850.

Zhou, H.; Liu, Q.; and Wang, Y. 2023. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10608–10617.

Zhou, L.; Zhou, Y.; Lam, T. L.; and Xu, Y. 2022. Context-aware mixture-of-experts for unbiased scene graph generation. *arXiv preprint arXiv:2208.07109*.